Investigation of suPAR as a Biomarker for Renal Dysfunction

Sarika Aggarwal, Vivian Garcia, Miriam Hu, and Paul Licht

July 6, 2018

Contents

Overview of the Clinical Issue

Since 2004, chronic kidney disease (CKD) has continued to affect approximately 14 percent of people in the United States. CKD is characterized by protein in the urine, high blood pressure, anemia, nerve damage, heart and blood vessel disease, and can eventually lead to kidney failure requiring dialysis or a kidney transplant in its later stages. CKD is defined by five stages based on a numeric value of glomerular filtration rate (GFR). However, in its earlier stages (1 to 3A), there are few noticeable symptoms which allows the disease to progress unnoticed until it is detrimental to a person's health. Furthermore, people that are diagnosed with CKD are generally known to have higher risk for comorbid conditions such as cardiovascular disease and diabetes which can complicate treatment.

If a biomarker can be identified to help detect CKD earlier, then treatment can be implemented to prevent the progression of kidney disease. In recent research, soluble urokinase-type plasminogen activator receptor (suPAR) has been investigated as a biomarker in the development of kidney disease, yet it is still unknown if increased suPAR levels can accurately determine future kidney dysfunction. Through this data analysis, we aimed to determine the relationship between suPAR and GFR as well as if suPAR levels can be used to predict decline in kidney function through CKD stage progression while considering covariates[1] ("Kidney Disease Statistics in the United States").

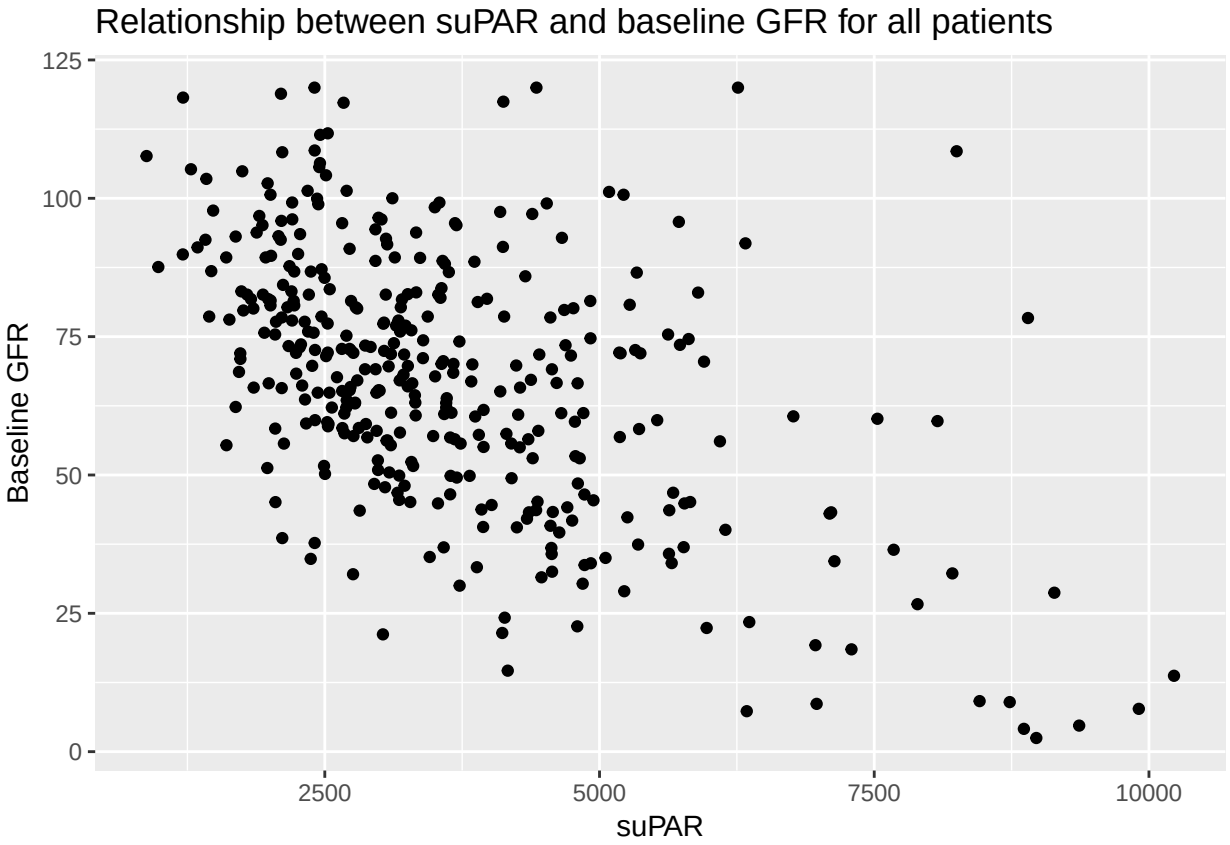
[1]Covariates: clinical CKD at baseline, C-reactive protein, age, sex, myocardial infarction, hypertension, diabetes, dyslipidemia, proteinuria, smoking, coronary artery disease (CAD), and use of angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor blockers (ARBs)

Summary of Patient Characteristics at Baseline

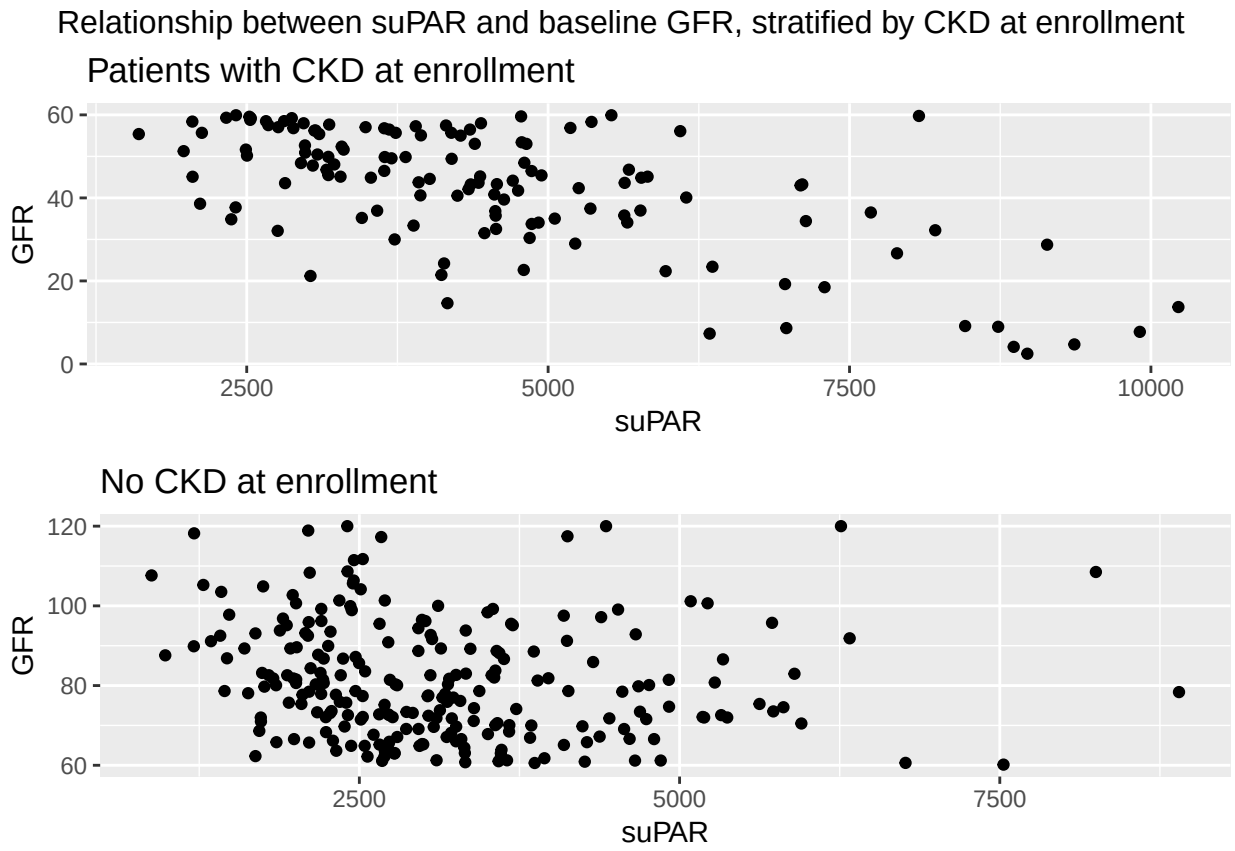
Table 1. Baseline Characteristics of Participants in Study		
Characteristic	**GFR≥60 (N=236)**	**GFR<60 (N=130)**
GFR at Baseline—mL/min/1.73m^2	81.1 ± 14.3	42.5 ± 14.5
suPAR—pg/mL	3138.5 ± 1265.8	4535.3 ± 1856.7
C-Reactive Protein (CRP)—mg/dL	5.0 ± 7.2	6.3 ± 11.5
Creatinine—mg/dL	1.0 ± 0.2	2.1 ± 2.4
Age—yrs	60.7 ± 12.3	67.6 ± 11.2
Sex–no. (%)		
Male	159 (67.4)	73 (56.2)
Female	77 (32.6)	57 (43.8)
Race—no. (%)		
Black	52 (22.0)	22 (16.9)
Non-Black	184 (78.0)	108 (83.1)
Body Mass Index (BMI)—kg/m^2	29.7 ± 6.2	29.4 ± 7.2
History of Myocardial Infarction—no. (%)		
Yes	59 (25.0)	34 (26.2)
No	173 (73.3)	92 (70.8)
Missing Information (N/A)	4 (1.7)	4 (3.1)
Diagnosis of Diabetes—no. (%)		
Yes	86 (36.4)	47 (36.2)
No	146 (61.9)	81 (62.3)
Missing Information (N/A)	4(1.7)	2 (1.5)
Hypertension—no. (%)		
Yes	166 (70.3)	99 (76.2)
No	66 (28.0)	28 (21.5)
Missing Information (N/A)	4(1.7)	3 (2.3)
Dyslipidemia—no. (%)		
Yes	154 (65.3)	89 (68.5)
No	78 (33.1)	37 (28.5)
Missing Information (N/A)	4(1.7)	4 (3.1)
Proteinuria—no. (%)		
Yes	9 (3.8)	13 (10)
No	227 (96.2)	117 (90)
History of Smoking—no. (%)		
Yes	129 (54.7)	70 (53.8)
No	99 (41.9)	58 (44.6)
Missing Information (N/A)	8(3.4)	2 (1.5)

Research Objectives

1. What is the relationship between suPAR and GFR at baseline?



To determine the relationship between suPAR and GFR at baseline, we ran a Pearson's correlation test. There is a moderately strong negative correlation between suPAR and baseline GFR, as evidenced by the correlation coefficient of -0.528. The correlation is also statistically significant with a very small p-value $< 2.2 \times 10^{-16}$. This means that as GFR increases, the amount of suPAR decreases, and vice versa.



We also stratified by whether the patient has CKD at enrollment (i.e. baseline GFR < 60). The relationship between suPAR and GFR changed after stratification. The scatterplots show a moderately strong negative correlation for patients with CKD, but for patients without CKD, the correlation is weak.

Among patients who had clinical CKD at enrollment ($n = 130$), the correlation coefficient is -0.655, which indicates a stronger negative correlation between suPAR and baseline GFR than we saw for all patients together. This association suggests that people with clinical CKD (and therefore lower GFR) will generally have higher amounts of suPAR.

Among patients who did *not* have clinical CKD at enrollment ($n = 236$), the correlation coefficient is -0.144, which indicates a much weaker correlation between suPAR and baseline GFR than we saw for all patients together. Baseline GFR decreases only slightly with increasing suPAR among patients without CKD.

2. Can suPAR levels at baseline predict decline of kidney function at one year?

We fitted a linear regression model for the difference in GFR (follow-up – baseline) against suPAR levels. Since a lower GFR score indicates worse kidney functioning, a positive difference indicates improvement, and a negative difference indicates worse functioning.

When suPAR level was the sole predictor in the model, it was statistically insignificant ($t = 1.81$, p-value $= 0.07$) and thus not a sufficient predictor for difference in GFR. However, in a multiple linear regression model using suPAR level, clinical CKD status as an indicator variable, and an interaction term between clinical CKD status and suPAR level as predictors of the difference in GFR, we found suPAR level ($t = -2.57$, p-value $= 0.01$) and the interaction between clinical CKD and suPAR level ($t = 2.63$, p-value < 0.01) to be statistically significant. Clinical CKD status ($t = 0.33$, p-value $= 0.74$) was not significant.[2]

[2]See Section 2 in the Appendix.

Baseline CKD Stage	Follow Up CKD Stage: Normal	Stage 1	Stage 2	Stage 3a	Stage 3b	Stage 4	Stage 5	Total
Normal	0	0	0	0	0	0	0	0
Stage 1	0	28	23	5	4	0	0	60
Stage 2	0	13	126	31	5	1	0	176
Stage 3a	0	1	25	24	16	0	0	66
Stage 3b	0	1	4	10	17	10	0	42
Stage 4	0	0	0	0	4	5	3	12
Stage 5	0	1	2	1	0	1	5	10
Total	0	44	180	71	46	17	8	366

Figure 1: Counts for baseline vs. follow-up CKD stages. Green indicates improvement, yellow indicates no change, and red indicates CKD stage progression.

13.05 percent of the variation in GFR difference is explained by the model with suPAR level, clinical CKD status, and the interaction between clinical CKD status and suPAR level. The fitted regression line is

$$\Delta\text{GFR} = -0.114 - 0.002(\text{suPAR}) + 1.555(\text{CKD status}) + 0.003(\text{suPAR*CKD status}).$$

The significance of the interaction term means that the relationship between suPAR and ΔGFR changes based on whether or not patients have clinical CKD at enrollment.

Difference in GFR *decreases* by 0.002 mL/min/1.73m^2 for each 1 pg/mL increase in suPAR when patients do *not* have clinical CKD at enrollment. Difference in GFR *increases* by 0.001 mL/min/1.73m^2 for each 1 pg/mL increase in suPAR when patients *do* have clinical CKD at enrollment.

So the fitted model when CKD = 0 is

$$\Delta\text{GFR} = -0.114 - 0.002(\text{suPAR}),$$

and the fitted model when CKD = 1 is

$$\Delta\text{GFR} = 1.441 + 0.001(\text{suPAR}).$$

3. Can suPAR measured at baseline predict the risk of CKD stage progression?

In order to evaluate CKD stage progression, we created a binary variable called progression which is equal to 1 if CKD stage progression has occurred (i.e. stage at follow-up is worse than at baseline) and 0 if it has not (i.e. stage at follow-up has improved or stayed the same as baseline).

We used a logistic regression model to determine whether suPAR measured at baseline can predict the risk of CKD stage progression. suPAR level is a continuous variable ranging from 878 pg/mL to 10,228 pg/mL.

The fitted odds are

$$\frac{\hat{p}}{1-\hat{p}} = \exp(-1.540 + (1.425 \times 10^{-4} * \text{suPAR})),$$

where $\hat{p}$ is the estimated probability that CKD progresses.

From the model, for every unit increase in suPAR, the odds of stage progression increases by a factor of 1.0001, which was found to be statistically significant (p-value = 0.0374). This means that for a one unit increase in suPAR, the odds of having CKD progression are roughly the same (very slight increase).

Although this increase in odds is statistically significant, it is probably too small to be clinically significant. We would need to know how much of an increase in suPAR is clinically meaningful. For example, maybe

a change of ±1 pg/mL of suPAR can be attributed to instrument error. If we consider 1,000 pg/mL to be the smallest clinically meaningful increase in suPAR, then for an increase of 1,000 pg/mL of suPAR, the odds of stage progression occuring increases by a factor of $(1.0001)^{1000} = 1.154$, which is large enough to be measured.

4. Are any categorical variables associated with CKD stage progression?

To determine which variables were associated with CKD stage progression, we ran a series of chi-square tests of independence for stage progression against each of the following categorical variables: clinical CKD at baseline, sex, history of myocardial infarction, coronary artery disease, hypertension, diabetes, dyslipidemia, and use of ACE inhibitors or ARBs[3]. We found that there was no association between any of the variables and CKD stage progression at a significance level of $\alpha = 0.05$.

	p-value	n
Myocardial infarction	0.801	358
Diabetes	0.214	360
Hypertension	0.126	359
Dyslipidemia	0.605	358
Proteinuria	0.195	366
Smoking	0.840	356
Coronary artery disease	0.910	350
Use of ACE/ARBs	0.999	363

All chi-square tests resulted in p-values larger than 0.05, so we fail to reject all null hypotheses that CKD stage progression is independent of each of the variables. There is not enough evidence to suggest that CKD stage progression and each of these categorical covariates are associated.

5. At baseline, what is the risk of clinical CKD for those with diabetes, hypertension, dyslipidemia, proteinuria, or coronary artery disease (CAD)? For smokers, users of ACE/ARBs? Do any associations change at follow-up?

By fitting a logistic regression model to the following potential risk factors[4] (diabetes, hypertension, dyslipidemia, proteinuria, history of smoking, coronary artery disease, and use of ACE inhibitors and/or angiotensin II receptor blockers) with CKD status as the outcome, we calculated the odds of increased CKD. The odds ratio between the presence of a risk factor at baseline and CKD was used to measure the risk association. The odds ratio represents the odds of CKD for an individual with a risk factor compared to an individual without the risk factor (e.g. an odds ratio of 1.4 for hypertension means that a patient with hypertension has 1.4 times the odds of having CKD as someone without hypertension and therefore a 40 percent increased risk relative to those without hypertension).

There are increased odds of CKD for patients who have hypertension (40.6 percent), dyslipidemia (21.9 percent), proteinuria (180 percent), or use of ACE inhibitors and/or angiotensin II receptor blockers (26.7 percent) as compared to patients who do not have these risk factors. However, the risk association between patients who have proteinuria and CKD is the only risk factor that is statistically significant (p-value = 0.02).

There are decreased odds of CKD for patients who have diabetes (1.5 percent), have ever smoked (7.4 percent), or have CAD (21.4 percent) as compared to patients who do not have these risk factors. However, none of these risk differences are statistically significant.

[3] All expected cell counts were > 5 for all chi-square tests.

[4] All risk factors are indicator (binary) variables.

Risk Factor for CKD	Odds Ratio BL	Odds Ratio FU
Diabetes	0.9851	1.5103
Hypertension	1.4057	1.5906
Dyslipidemia	1.2185	1.4169
Proteinuria	2.8025	2.4075
Ever Smoked	0.9262	1.01
Coronary Artery Disease	0.7863	0.8427
Use of ACE inhibitors	1.2668	1.3547

Figure 2: Comparison of odds ratios of having CKD given different risk factors at baseline and follow-up.

At follow-up, the significance of each risk association remained the same at $\alpha = 0.05$. Proteinuria was again the only statistically significant risk factor. However, there was a marked difference in diabetes, as its odds ratio represented an increase in risk rather than a decrease. Also, the odds ratio for smoking changed from a slight decrease to roughly no change in risk. Every other risk factor had only slight changes in the magnitude of risk.

Code

```
library(tidyverse) # data cleaning and graphs
library(broom) # data cleaning
library(readxl) # read an Excel file
library(GGally) # graphs
library(gridExtra) # graphs
library(lmtest) # diagnostic tests
library(e1071) # diagnostic tests

GFR <- read_excel("SIBS_GFRdata.xlsx", na = ".") # read in the data

# Make indicator for whether patient has CKD at enrollment
GFR <- mutate(GFR, BL_CKD = ifelse(BL_GFR < 60, 1, 0))
GFR <- mutate(GFR, FU_CKD = ifelse(BL_GFR < 60, 1, 0))

# Make indicator for whether patient's CKD worsened at follow-up:
# Step 1: Recode baseline CKD stages as numbers
GFR <- mutate(GFR, BL_num_CKD = BL_GFR)
GFR$BL_num_CKD <- ifelse(GFR$BL_num_CKD >= 120, 0,
                  ifelse(GFR$BL_num_CKD >= 90, 1,
                  ifelse(GFR$BL_num_CKD >= 60, 2,
                  ifelse(GFR$BL_num_CKD >= 45, 3,
                  ifelse(GFR$BL_num_CKD >= 30, 4,
                  ifelse(GFR$BL_num_CKD >= 15, 5, 6))))))

# Step 2: Recode follow-up CKD stages as numbers
GFR <- mutate(GFR, FU_num_CKD = FU_GFR)
GFR$FU_num_CKD <- ifelse(GFR$FU_num_CKD >= 120, 0,
                  ifelse(GFR$FU_num_CKD >= 90, 1,
                  ifelse(GFR$FU_num_CKD >= 60, 2,
```

```
                  ifelse(GFR$FU_num_CKD >= 45, 3,
                  ifelse(GFR$FU_num_CKD >= 30, 4,
                  ifelse(GFR$FU_num_CKD >= 15, 5, 6))))))

# Step 3: Create the new variable by subtracting baseline from follow-up
GFR <- mutate(GFR, progression = FU_num_CKD - BL_num_CKD) # if positive, CKD got worse
GFR$progression <- ifelse(GFR$progression > 0, 1, 0) # 1 for worse; 0 for same or better
```

```
# Pearson's correlation for suPAR and baseline GFR (all patients):
cor.test(GFR$suPARpgml, GFR$BL_GFR)

# Spearman's correlation for suPAR and baseline GFR (all patients):
cor.test(GFR$suPARpgml, GFR$BL_GFR, method = "spearman") # agrees with Pearson

# Pearson's correlation among those who have CKD at enrollment:
cor.test(GFR$suPARpgml[GFR$BL_CKD == 1], GFR$BL_GFR[GFR$BL_CKD == 1])

# Spearman's correlation among those who have CKD at enrollment:
cor.test(GFR$suPARpgml[GFR$BL_CKD == 1], GFR$BL_GFR[GFR$BL_CKD == 1],
         method = "spearman") # agrees with Pearson

# Pearson's correlation among those who do not have CKD at enrollment:
cor.test(GFR$suPARpgml[GFR$BL_CKD == 0], GFR$BL_GFR[GFR$BL_CKD == 0])

# Spearman's correlation among those who have CKD at enrollment:
cor.test(GFR$suPARpgml[GFR$BL_CKD == 0], GFR$BL_GFR[GFR$BL_CKD == 0],
         method = "spearman") # agrees with Pearson
```

```
# Create a new variable for the difference in GFR (follow-up - baseline):
GFR$diff_GFR <- GFR$FU_GFR - GFR$BL_GFR # positive indicates improvement

# Simple linear regression model with suPAR as the only predictor:
mod2 <- lm(diff_GFR ~ suPARpgml, data = GFR)

# Multiple linear regression model with suPAR, baseline CKD, and interaction term:
mod2.inter <- lm(diff_GFR ~ suPARpgml + BL_CKD + suPARpgml*BL_CKD, data = GFR)
```

```
# Logistic regression model for CKD stage progression vs. suPAR:
log1 <- glm(progression ~ suPARpgml, data = GFR, family = "binomial")
exp(coef(log1)) # odds ratio
```

```
# Chi-square tests of independence for each categorical covariate and CKD stage progression
chi.mi <- chisq.test(GFR$EverMI, GFR$progression)
chi.diabetes <- chisq.test(GFR$DM, GFR$progression)
chi.hypertension <- chisq.test(GFR$HTN, GFR$progression)
chi.dyslipidemia <- chisq.test(GFR$Dyslipidemia, GFR$progression)
chi.proteinuria <- chisq.test(GFR$Proteinuria, GFR$progression)
chi.smokers <- chisq.test(GFR$EverSmoked, GFR$progression)
chi.cad <- chisq.test(GFR$CAD, GFR$progression)
chi.ace <- chisq.test(GFR$acearb, GFR$progression)

# Print chi-squared test statistic and p-value for all of the above:
chi.mi
chi.diabetes
```

```
chi.hypertension
chi.dyslipidemia
chi.proteinuria
chi.smokers
chi.cad
chi.ace
```

```
# Logistic regression models for CKD status vs. each categorical covariate at baseline:
diabetes <- glm(BL_CKD ~ DM, data = GFR, family = "binomial")
hypertension <- glm(BL_CKD ~ HTN, data = GFR, family = "binomial")
dyslipidemia <- glm(BL_CKD ~ Dyslipidemia, data = GFR, family = "binomial")
proteinuria <- glm(BL_CKD ~ Proteinuria, data = GFR, family = "binomial")
smokers <- glm(BL_CKD ~ EverSmoked, data = GFR, family = "binomial")
heartdisease <- glm(BL_CKD ~ CAD, data = GFR, family = "binomial")
ACE <- glm(BL_CKD ~ acearb, data = GFR, family = "binomial")

# Summary output (parameter estimates, p-values, etc.) for all of the above:
summary(diabetes)
summary(hypertension)
summary(dyslipidemia)
summary(proteinuria)
summary(smokers)
summary(heartdisease)
summary(ACE)

# Logistic regression models for CKD status vs. each categorical covariate at follow-up:
FU_diabetes <- glm(FU_CKD ~ DM, data = GFR, family = "binomial")
FU_hypertension <- glm(FU_CKD ~ HTN, data = GFR, family = "binomial")
FU_dyslipidemia <- glm(FU_CKD ~ Dyslipidemia, data = GFR, family = "binomial")
FU_proteinuria <- glm(FU_CKD ~ Proteinuria, data = GFR, family = "binomial")
FU_smoker <- glm(FU_CKD ~ EverSmoked, data = GFR, family = "binomial")
FU_heartdisease <- glm(FU_CKD ~ CAD, data = GFR, family = "binomial")
FU_ACE <- glm(FU_CKD ~ acearb, data = GFR, family = "binomial")

# Summary output for all of the above:
summary(FU_diabetes)
summary(FU_hypertension)
summary(FU_dyslipidemia)
summary(FU_proteinuria)
summary(FU_smokers)
summary(FU_heartdisease)
summary(FU_ACE)
```

Summary and Future Research

Objective 1

- Moderately strong negative correlation between suPAR levels and GFR at baseline (all patients) ($r = -0.528$)
- Moderately strong negative correlation between suPAR levels and GFR at baseline (patients with clinical CKD at baseline) ($r = -0.655$)

- Weak negative correlation between suPAR levels and GFR at baseline (patients without clinical CKD at baseline) ($r = -0.144$)

Objective 2

- SuPAR level does not provide enough information to be the sole predictor of GFR at baseline ($p = 0.07$).
- The interaction term between suPAR levels and CKD status should be included in the linear regression model/

Objective 3

- For a 1 pg/mL increase in suPAR, the odds of having CKD progression increase *very* slightly but are essentially the same at the 1 pg/mL level.

Objective 4

- None of the categorical variables have an association with CKD stage progression.

Objective 5

- Increased odds of CKD if a person has hypertension, dyslipidemia, proteinuria, or use of ACE inhibitors and/or angiotensin II receptor blockers
- Decreased odds in CKD for patients who have diabetes, have smoked before, or have CAD
- Changes of odds from baseline to follow-up: diabetes (decreased to increased odds of CKD), smoking (decreased to approximately equal odds of CKD)

Percentage of Participants who Progressed by Stage

	Better	Worse	No Change
Normal	N/A	N/A	N/A
Stage 1	0	53.3	46.7
Stage 2	7.4	21	71.6
Stage 3a	39.4	24.2	36.4
Stage 3b	35.7	23.8	40.5
Stage 4	33.3	25	41.2
Stage 5	50	0	50

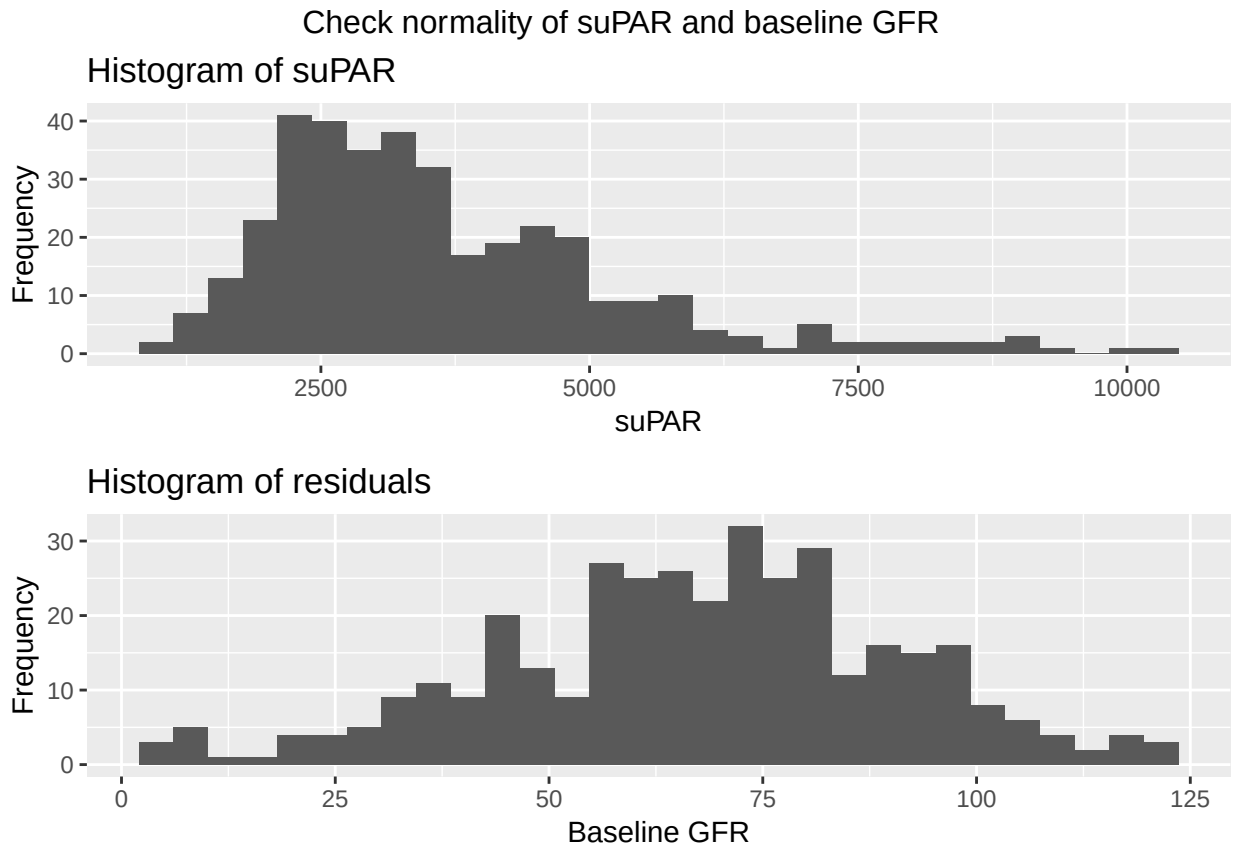
Figure 3: Percent of patients' status at follow-up by initial stage of CKD.

Additional Questions

- According to the data, more people saw improvements in CKD at later stages (3A and onwards) rather than at earlier stages (1 and 2). Further, at Stage 3, there are more health complications, so people are more likely to seek out healthcare and make lifestyle changes.
  - Were these patients treated? Could this be why more patients saw improvements in CKD stage at higher stages (That is, Stage 1 and Stage 2 are not taken as seriously). Does the progression of the disease increase more rapidly once a person is diagnosed with clinical CKD?
- According to the National Institute of Diabetes and Digestive and Kidney Disease, people diagnosed with CKD are generally known to have higher risk for cardiovascular disease and diabetes. Our data does not show a significance. Is there a known reason why?
- What magnitude of an increase or decrease in suPAR levels (pg/mL) is considered clinically meaningful?

Appendix

Assumptions for Objective 1



The scatterplots shown in Objective 1 show that the linearity assumption is met.

From the histograms, we were concerned that the normality assumption for Pearson's correlation would be violated because the distribution of suPAR is right-skewed. However, Pearson's and Spearman's correlation agree for every test, so we will report Pearson's correlation for this research question.

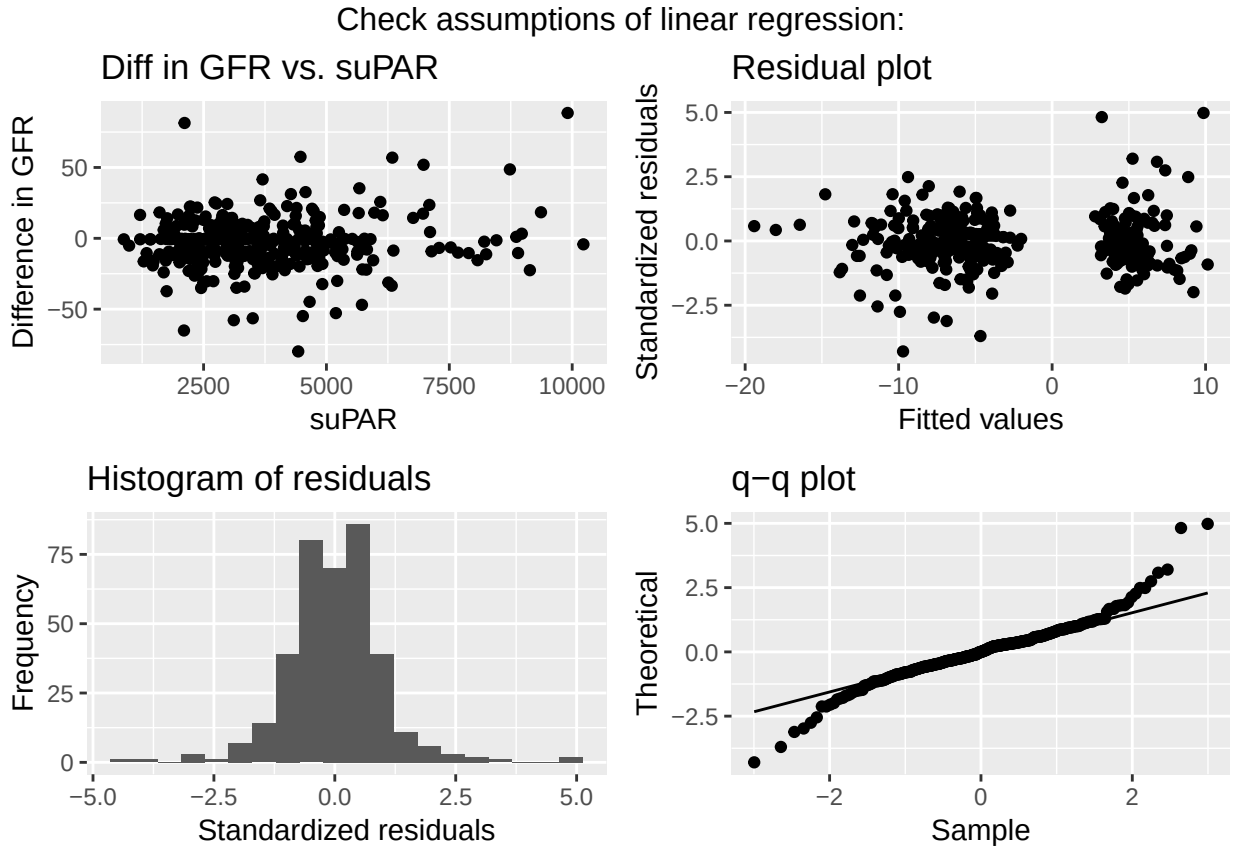
Output for Objective 2

Table 2: Simple linear regression model with suPAR as the only predictor for difference in GFR

term	estimate	std.error	statistic	p.value
(Intercept)	-6.241	2.218	-2.815	0.005
suPARpgml	0.001	0.001	1.812	0.071

Table 3: Multiple linear regression model with suPAR, baseline CKD, and interaction term

term	estimate	std.error	statistic	p.value
(Intercept)	-0.114	2.858	-0.040	0.968
suPARpgml	-0.002	0.001	-2.567	0.011
BL_CKDYes	1.555	4.760	0.327	0.744
suPARpgml:BL_CKDYes	0.003	0.001	2.630	0.009



From examining the histogram, the linearity assumption is met. The residual plot shows unequal variance, and this is confirmed by the Breusch-Pagan test for heteroscedasticity (p-value = 0.006). We reject the null hypothesis that the residuals have constant variance. This is probably because we do not have much data for patients whose GFR decreased greatly, so their variance is small. However, there does not appear to be a funnel shape in the residual plot, so we will proceed with caution.

We also need to check to make sure the residuals are normally distributed:

- **Unimodal:** Yes, the distribution of the standardized residuals appears unimodal, according to the histogram. We can see a slight dip near the center, but it's not large enough to cause concern.
- **Mean:** 0.00038
- **Median:** −0.0011
- **Skewness:** 0.30
  - This indicates that the distribution of the standardized residuals is approximately symmetric, which is supported by the mean and median being extremely close.
- **Kurtosis:** 4.23
  - This indicates that the distribution of the standardized residuals is heavy-tailed.

- **Shapiro-Wilk test:** p-value $= 1.40 \times 10^{-10}$
  - Since the p-value is very small, we would reject the null hypothesis that the standardized residuals are normally distributed. However, the Shapiro-Wilk test is extremely sensitive when the number of observations is large. Instead, we examine the q-q plot, which looks straight except for some deviation in the tails. Since linear regression is robust to non-normality, this is not a problem.